INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team MESCAL

Middleware Efficiently SCALable

Grenoble - Rhône-Alpes

THEME NUM

Table of contents

The MESCAL project-team is a common project-team supported by CNRS, INPG, UJF and INRIA located in the LIG laboratory (UMR 5217).

1. Team

Research Scientist

Bruno Gaujal [Research Director (DR) INRIA, HdR]
Corinne Touati [Research Associate (CR) INRIA]
Derrick Kondo [Research Associate (CR) INRIA]
Arnaud Legrand [Research Associate (CR) CNRS]

Faculty Member

Yves Denneulin [Professor, Grenoble INP, HdR]
Brigitte Plateau [Professor, Grenoble INP, HdR]
Vania Martin [Associate Professor, UJF]
Jean-François Méhaut [Professor, UJF, HdR]
Florence Perronnin [Associate Professor,UJF]
Olivier Richard [Associate Professo UJF, INRIA Delegation]
Jean-Marc Vincent [Associate Professor, UJF]
Jean-Michel Fourneau [Professor, HdR]

Technical Staff

Joseph Emeras [Engineer Assistant]
Kiril Georgiev [Engineer Assistant]

PhD Student

Hamza Adamou [2007, University of Yaoundé, Cameroon]
Carlos Jaime Barrios Hernandez [2005, EGIDE, co-tutelle]
Rémi Bertin [2007, ANR DOCCA]
Léonardo Brenner [2004, Brazilian CAPES scholarship]
Marcia Cristina Cera [2008, Brazilian CAPES scholarship, cotutelle]
Rodrigue Chakode Noumowe [2008, Minalogic CILOE scholarship]
Pierre Coucheney [2008, INRIA-Alcatel Lucent scholarship]
Nicolas Gast [2007, AC]
Yiannis Georgiou [2006, CIFRE BULL scholarship]
Ahmed Harbaoui [2006, CIFRE France Télécom R&D scholarship]
Hussein Joumma [2006, MNRT scholarship]
Lucas Nussbaum [2005, BDI-CNRS MNRT scholarship]
Matthieu Ospici [2008, CIFRE BULL scholarship]
Carlos Prada Rojas [2007, CIFRE STMicroelectronics]
Christiane Ribeiro [2008, Brazilian CAPES scholarship, cotutelle]
Afonso Sales [2005, Brazilian CAPES scholarship]
Nazha Touati [2004, RhÃ´ne-Alpes scholarship]
Pedro Antonio Velho [2006, Brazilian CAPES scholarship]
Jérome Vienne [2006, CIFRE BULL scholarship]
Brice Videau [2005, MNRT scholarship]
Blaise Yenké [2004, Ngaundere University scholarship]

Post-Doctoral Fellow

Ana Bušić [ANR SMS, October 2007 - July 2008]
Bahman Javadi Jahentigh [November 2008]

Administrative Assistant

Ahlem Zammit-Boubaker [Secretary (SAR) INRIA]

2. Overall Objectives

2.1. Presentation

MESCAL is a project-team of INRIA jointly with UJF and INPG universities and CNRS, created in 2005 as an offspring of the former APACHE project-team, together with MOAIS.

MESCAL's research progress and objective were evaluated by INRIA in 2008. The MESCAL project-team received positive evaluations and useful feedback. As such, the project-team was extended for another 4 years by the INRIA evaluation commission.

2.2. Objectives

The recent evolutions in computer networks technology, as well as their diversification, goes with a tremendous change in the use of these networks: applications and systems can now be designed at a much larger scale than before. This scaling evolution concerns at the same time the amount of data, the number of computers, the number of users, and the geographical diversity of these users.

This race towards *large scale* computing questions many assumptions underlying parallel and distributed algorithms and operating middleware. Today, most software tools developed for average size systems cannot be run on large scale systems without a significant degradation of their performances.

The goal of the MESCAL project-team is to design and validate efficient exploitation mechanisms (middleware and system services) for large distributed infrastructures.

MESCAL's target applications are intensive scientific computations such as cellular micro-physiology, protein conformations, particle detection, combinatorial optimization, Monte Carlo simulations, and others. Such applications are constituted of a large set of independent, equal-sized tasks and therefore may benefit from large-scale computing platforms. Initially executed on large dedicated clusters (CRAY, IBM, COMPAQ), they have been recently deployed on collections of homogeneous clusters aggregating a large number of commodity components. The experience showed that such clusters offer a huge computing power at a very reasonable price. MESCAL's target infrastructures are aggregations of commodity components and/or commodity clusters at metropolitan, national or international scale. Examples of target infrastructures are grids obtained through sharing of available resources inside autonomous computing services, lightweight grids (such as the local CIMENT Grid) which are limited to trusted autonomous systems, clusters of intranet resources (Condor) or aggregation of Internet resources (SETI@home, XtremWeb).

MESCAL's methodology in order to ensure **efficiency** and **scalability** of proposed mechanisms is based on systematic modeling and performance evaluation of target architectures, software layers and applications.

3. Scientific Foundations

3.1. Large System Modeling and Analysis

Keywords: *Discrete event dynamic systems*, *Markov chains*, *Performance evaluation*, *Petri nets*, *Queuing networks*, *Simulation*.

Participants: Bruno Gaujal, Derrick Kondo, Arnaud Legrand, Florence Perronnin, Brigitte Plateau, Olivier Richard, Corinne Touati, Jean-Marc Vincent.

Understanding qualitative and quantitative properties of distributed systems and parallel applications is a major issue. The *a posteriori* analysis of the behavior of the system or the design of predictive models are notoriously challenging problems.

Indeed, large distributed systems contain many different features (processes, threads, jobs, messages, packets) with intricate interactions between them (communications, synchronizations). The analysis of the global behavior of the system requires to take into account large data sets.

As for *a priori* models, our current research focuses on capturing the distributed behavior of large dynamic architectures. Actually, both formal models and numerical tools are being used to get predictions on the behavior of large systems.

For large parallel systems, the non-determinism of parallel composition, the unpredictability of execution times and the influence of the outside world are usually expressed in the form of multidimensional stochastic processes which are continuous in time with a discrete state space. The state space is often infinite or very large and several specific techniques have been developed to deal with what is often termed as the "curse of dimensionality".

MESCAL deals with this problem using several complementary tracks:

- Behavior analysis of highly distributed systems,
- Simulation algorithms able to deal with very large systems,
- Fluid limits (used for simulation and analysis),
- Decomposition of the state space,
- Structural and qualitative analysis,
- Game theory methods for resolving resource contention.

3.1.1. Behavior analysis of highly distributed systems

The development of highly distributed architectures running widely spread applications requires to elaborate new methodologies to analyze the behavior of systems. Indeed, runtime systems on such architectures are empirically tuned. Analysis of executions are generally manually performed on *post-mortem* traces that have been extracted with very specific tools. This tedious methodology is generally motivated by the difficulty to characterize the resources of such systems. For example, big clusters, grids or peer-to-peer (P2P) [1] networks present properties of size, heterogeneity, dynamicity that are usually not taken into account in classical system models. The asynchrony of the architecture also induces perturbations in the behavior of the application leading to significant slow-down that should be avoided. Therefore, when defining the workload of the system, the distributed nature of applications should be taken into account with a specific focus on problems related to synchronizations.

3.1.2. Simulation of distributed systems

Since the advent of distributed computer systems, an active field of research has been the investigation of *scheduling* strategies for parallel applications. The common approach is to employ scheduling heuristics that approximate an optimal schedule. Unfortunately, it is often impossible to obtain analytical results to compare the efficiency of these heuristics. One possibility is to conduct large numbers of back-to-back experiments on real platforms. While this is possible on tightly-coupled platforms, it is infeasible on modern distributed platforms (i.e. Grids or peer-to-peer environments) as it is labor-intensive and does not enable repeatable results. The solution is to resort to *simulations*. Simulations not only enable repeatable results but also make it possible to explore wide ranges of platform and application scenarios.

The SIMGRID framework enables the simulation of distributed applications in distributed computing environments for the specific purpose of developing and evaluating scheduling algorithms. This software is the result of a long-time collaboration with Henri CASANOVA (University of California, San Diego).

3.1.3. Perfect Simulation

Using a constructive representation of a Markovian queuing network based on events (often called GSMPs), we have designed a perfect simulation tool computing samples distributed according to the stationary distribution of the Markov process with no bias. Two softwares have been developed. ψ analyzes a Markov chain using its transition matrix and provides perfect samples of cost functions of the stationary state. ψ^2 samples the stationary measure of Markov processes using directly the queuing network description. Some monotone networks with up to 10^{50} states can be handled within minutes over a regular PC.

[1] Our definition of peer-to-peer is a network (mainly the Internet) over which a large number of autonomous entities contribute to the execution of a single task.

3.1.4. Fluid models

When the size of systems grows very large, one may use asymptotic techniques to get a faithful estimate of their behaviors. One such tools is mean field analysis and fluid limits, that are used on a modeling and simulation level. One recent significant application is call centers. Another one is peer to peer systems. Web caches as well as peer-to-peer systems must be able to serve a set of customers which is both large (several tens of thousands) and highly volatile (with short connection times). These features make analysis difficult when classical approaches (like Markovian Models or simulation) are used. We have designed simple fluid models to get rid of one dimension of the problem. This approach has been applied to several systems of web caches (such as Squirrel) and to peer-to-peer systems (such as BitTorrent). This helps to get a better understanding of the behavior of the system and to solve several optimization problems.

3.1.5. Markov Chain Decomposition

The first class of models we will be using is Continuous time Markov chains (CTMC). The usefulness of Markov models is undisputed, as attested by the large number of modeling tools implementing Markov solvers. However their practical applications are limited by the *state-space explosion* problem, which puts excessive demands on memory and execution time when studying large real-life systems. Continuous-time Stochastic Automata Networks describe a system as a set of subsystems that interact. Each subsystem is modeled by a stochastic automaton, and some rules between the states of each automaton describe the interactions between subsystems. The main challenge is to come up with ways to compute the asymptotic (or transient) behavior of the system without ever generating the whole state space. Several techniques have been developed in our group based on bounds, lumpability, symmetry and properties of the Kronecker product. Most of them have been integrated in a software tool (PEPS) which is openly available.

3.1.6. Discrete Event Systems

The interaction of several processes through synchronization, competition or superposition within a distributed system is a big source of difficulties because it induces a state space explosion and a non-linear dynamic behavior. The use of exotic algebra, such as (min,max,plus) can help. Highly synchronous systems become linear in this framework and therefore are amenable to formal solutions. More complicated systems are neither linear in (max,plus) nor in the classical algebra. Several qualitative properties have been established for a large class of such systems called free-choice Petri nets (sub-additivity, monotonicity or convexity properties). Such qualitative properties are sometimes enough to assess the class of routing policies optimizing the global behavior of the system. They are also useful to design efficient numerical tools computing their asymptotic behavior.

3.1.7. Game Theory Methods for Resolving Resource Contention

Resources in large-scale distributed platforms (Grid computing platforms, enterprise networks, peer-to-peer systems) are shared by a number of users having conflicting interests who are thus prone to act selfishly. A natural framework for studying such non-cooperative individual decision-making is game theory. In particular, game theory models the decentralized nature of decision-making.

It is well known that such non-cooperative behaviors can lead to important inefficiencies and unfairness. In other words, individual optimizations often results in global resource waste. In the context of game theory, a situation in which all users selfishly optimize their own utility is known as a *Nash equilibrium* or *Wardrop equilibrium*. In such equilibria, no user has interest in unilaterally deviating from its strategy. Such policies are thus very easy to implement in a fully distributed system and have some stability properties. However, a possible consequence is the *Braess paradox* in which the increase of resource happens at the expense of *every* user. This is why, the study of the occurrence and degree of such inefficiency is of crucial interest. Up until now, little is known about general conditions for optimality or degree of efficiency of these equilibria, in a general setting.

Many techniques have been developed to enforce some form of collaboration and improve these equilibria. In this context, it is generally prohibitive to take joint decisions so that a global optimization cannot be achieved. A possible option relies on the establishment of virtual prices, also called *shadow prices* in congestion networks. These prices ensure a rational use of resources. Equilibria can also be improved by advising policies to mobiles such that any user that does not follow these pieces of advice will necessarily penalize herself (*correlated equilibria*).

3.2. Management of Large Architectures

Keywords: *Administration*, *Clusters*, *Deployment*, *Grids*, *Job scheduler*, *Peer-to-peer*.

Participants: Derrick Kondo, Arnaud Legrand, Olivier Richard, Corinne Touati, Vania Marangozova.

Most distributed systems deployed nowadays are characterized by a high dynamism of their entities (participants can join and leave at will), a potential instability of the large scale networks (on which concurrent applications are running), and the increasing probability of failure. Therefore, as the size of the system increases, it becomes necessary that it adapts automatically to the changes of its components, requiring a self-organization of the system with respect to the arrival and departure of participants, data, or resources.

As a consequence, it becomes crucial to understand and model the behavior of large scale systems, to efficiently exploit these infrastructures. In particular it is essential to design dedicated algorithms and infrastructures handling a large amount of users and/or data.

MESCAL deals with this problem using several complementary tracks:

- Fairness in large-scale distributed systems,
- Deployment and management tools,
- Scalable batch scheduler for clusters and grids.

3.2.1. Fairness in large-scale distributed systems

Large-scale distributed platforms (Grid computing platforms, enterprise networks, peer-to-peer systems) result from the collaboration of many people. Thus, the scaling evolution we are facing is not only dealing with the amount of data and the number of computers but also with the number of users and the diversity of their behavior. In a high-performance computing framework, the rationale behind this joining of forces is that most users need a larger amount of resources than what they have on their own. Some only need these resources for a limited amount of time. On the opposite some others need as many resources as possible but do not have particular deadlines. Some may have mainly tightly-coupled applications while some others may have mostly embarrassingly parallel applications. The variety of user profiles makes resources sharing a challenge. However resources have to be *fairly* shared between users, otherwise users will leave the group and join another one. Large-scale systems therefore have a real need for fairness and this notion is missing from classical scheduling models.

3.2.2. Tools to operate clusters

The MESCAL project-team studies and develops a set of tools designed to help the installation and the use of a cluster of PCs. The first version had been developed for the icluster1 platform exploitation. The main tools are a scalable tool for cloning nodes (KA-DEPLOY) and a parallel launcher based on the TAKTUK project (now developed by the MOAIS project-team). Many interesting issues have been raised by the use of the first versions among which we can mention environment deployment, robustness and batch scheduler integration. A second generation of these tools is thus under development to meet these requirements.

The new KA-DEPLOY has been retained as the primary deployment tool for the experimental national grid GRID'5000.

3.2.3. Simple and scalable batch scheduler for clusters and grids

Most known batch schedulers (PBS, LSF, Condor, ...) are of old-fashioned conception, built monolithically, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150000 lines of code for OpenPBS), offering a growing number of functions that are, most of the time, not used. In such a context, it becomes hard to control both the robustness and the scalability of the whole system.

OAR is an attempt to address these issues. Firstly, OAR is written in a very high level language (Perl) and makes intensive use of high level tools (MySql and TAKTUK), thereby resulting in a concise code (around 5000 lines of code) easy to maintain and extend. This small code as well as the choice of widespread tools (MySql) are essential elements that ensure a strong robustness of the system. Secondly, OAR makes use of SQL requests to perform most of its job management tasks thereby getting advantage of the strong scalability of most database management tools. Such scalability is further improved in OAR by making use of TAKTUK to manage nodes themselves.

3.3. Migration and resilience

Keywords: *Fault tolerance*, *distributed algorithms*, *migration*.

Participants: Yves Denneulin, Jean-François Méhaut.

Making a distributed system reliable has been and remains an active research domain. Nonetheless this has not so far lead to results usable in an intranet or federal architecture for computing. Most propositions address only a given application or service. This may be due to the fact that until clusters and intranet architectures arose, it was obvious that client and server nodes were independent. So, a fault or a predictable disconnection on most of the nodes didn't lead to a complete failure of the system. This is not the case in parallel scientific computing where a fault on a node can lead to a data loss on thousands of other nodes. The reliability of the system is hence a crucial point. MESCAL's work on this topic is based on the idea that each process in a parallel application will be executed by a group of nodes instead of a single node: when the node in charge of a process fails, another in the same group can replace it in a transparent way for the application.

There are two main problems to be solved in order to achieve this objective. The first one is the ability to migrate processes of a parallel, and thus communicating, application without enforcing modifications. The second one is the ability to maintain a group structure in a completely distributed way. The first one relies on a close interaction with the underlying operating systems and networks, since processes can be migrated in the middle of a communication. This can only be done by knowing how to save and replay later all ongoing communications, independently of the communications. Freezing a process to restore it on another node is also an operation that requires collaboration of the operating system and a good knowledge of its internals. The other main problem (keeping a group structure) belongs to the distributed algorithms domain and is of a much higher level nature.

Future work will concern the behavior analysis of checkpoint systems in order to predict precisely critical operations to optimize resource usage (network and disk bandwidth).

3.4. Large scale data management

Keywords: *Fault tolerance*, *distributed algorithms*, *migration*.

Participants: Yves Denneulin, Vania Marangozova.

In order to use large data, it is necessary (but not always sufficient, as seen later) to efficiently store and transfer them to a given site (a set of nodes) where it is going to be used. The first step toward this achievement is the construction of a file system that is an extension of NFS for the grid environment. The second step is an efficient transfer tool that provides throughput close to optimal (*i.e.* the capacity of the underlying hardware).

3.4.1. Fast distributed storage over a cluster

Our goal here is to design a distributed file system for clusters that enables one to store data over a set of nodes (instead of a single one). It was designed to permit the usage of a set of disks to optimize memory allocations. It is important for performance and simplicity that this new file system has little overhead for access and updates. From a user point of view, it is used just as a classical NFS. From the server point of view, however, the storage is distributed over several nodes (possibly including the users).

The mounting point is only in charge of the meta-data, name, owner, access permissions, size, inodes, and etc., of the files while their content is stored on separate nodes. Every read or write request is received by the meta-server, the mounting point, which sends them to the relevant storage nodes, called IOD for Input/Output Daemon which will serve the request and send the result to the client.

Two implementations were done, one at the user level and one at the kernel level. Performances are good for read operations, for example 150MBs/sec for 16 IODs connected through a 100Mb/s for 16 clients. For write operations performances are limited by the bandwidth available for the meta-server which is a significant bottleneck.

3.4.2. Reliable distribution of data

Storage distribution on a large set of disks raises the reliability problem: more disks mean a higher fault rate. To address this problem we introduced in NFSP a redundancy on the IODs, the storage nodes by defining VIOD, Virtual IOD, which is a set of IODs that contain exactly the same data. So when an IOD fails another one can serve the same data and continuity of service is insured though. This doesn't modify the way the file-system is used by the clients: distribution and replication remain transparent. Several consistency protocols are proposed with various levels of performance; they all enforce at least the NFS consistency which is expected by the client.

4. Application Domains

4.1. Introduction

Applications in the fields of numerical simulation, image synthesis, and processing are typical of the user demand for high performance computing. In order to confront our proposed solutions for parallel computing with real applications, the project-team is involved in collaborations with end-users to help them parallelize their applications.

4.2. On-demand Geographical Maps

Participant: Jean-Marc Vincent.

This joint work involves the UMR 8504 Géographie-Cité, LSR-IMAG, UMS RIATE and the Maisons de l'Homme et de la Société.

Improvements in the Web developments have opened new perspectives in interactive cartography. Nevertheless existing architectures have some problems to perform spatial analysis methods that require complex calculus over large data sets. Such a situation involves some limitations in the query capabilities and analysis methods proposed to users. The HyperCarte consortium with LSR-IMAG, Géographie-cité and UMR RIATE proposes innovative solutions to these problems. Our approach deals with various areas such as spatio-temporal modeling, parallel computing and cartographic visualization that are related to spatial organizations of social phenomena.

Nowadays, analysis are done on huge heterogeneous data set. For example, demographic data sets at nuts 5 level, represent more than 100.000 territorial units with 40 social attributes. Many algorithms of spatial analysis, in particular potential analysis are quadratic in the size of the data set. Then adapted methods are needed to provide "user real time" analysis tools.

4.3. Seismic simulations

Participant: Jean-François Méhaut.

Numerical modeling of seismic wave propagation in complex three-dimensional media is an important research topic in seismology. Several approaches will be studied, and their suitability with respect to the specific constraints of NUMA architectures shall be evaluated. These modeling approaches will rely on modern numerical schemes such as spectral elements, high-order finite differences or finite elements applied to realistic 3D models. The NUMASIS project (see Section 8.2.4) will focus on issues related to parallel algorithms (distribution, scheduling) in order to optimize computations based on such numerical schemes by taking advantage of execution frameworks developed for NUMA architectures.

These approaches will be tested and validated on applications related to seismic risk assessment. Recent seismic events as those in Asia have evidenced the crucial research and development needs in this field. Some regions in France may as well be prone to such risks (French Riviera, Alps, French Antilles,...) and the experiments in the NUMASIS project will be carried out using some of the available data from these regions.

4.4. The CIMENT project

Participant: Olivier Richard.

The CIMENT project (Intensive Computing, Numerical Modeling and Technical Experiments, http://ciment.ujf-grenoble.fr/) gathers a wide scientific community involved in numerical modeling and computing (from numerical physics and chemistry to astrophysics, mechanics, bio-modeling and imaging) and the distributed computer science teams from Grenoble. Among these various application domains, there is a huge demand to manage executions of large sets of independent jobs. These sets have between 10,000 to 100,000 jobs each. Providing a middleware able to steer such an amount of jobs is a challenge. The CiGri middleware project addresses this issue in a grid infrastructure.

The aim of the CiGri project is to gather the unused computing resource from intranet infrastructure and to make it available for large scale applications. This grid is based on two software tools. The CiGri server software is based on a database and offers a user interface for launching grid computations (scripts and web tools). It interacts with the computing clusters through a batch scheduler software. CiGri is compatible with classical batch systems like PBS, but an efficient batch software (OAR, http://oar.imag.fr/) has been developed by the MESCAL and MOAIS project-teams for the easy integration and testing of scheduling tools.

5. Software

5.1. Tools for cluster management and software development

The large-sized clusters and grids show serious limitations in many basic system softwares. Indeed, the launching of a parallel application is a slow and significant operation in heterogeneous configurations. The broadcast of data and executable files is widely under the control of users. Available tools do not scale because they are implemented in a sequential way. They are mainly based on a single sequence of commands applied over all the cluster nodes. In order to reach a high level of scalability, we propose a new design approach based on a parallel execution. We have implemented a parallelization technique based on spanning trees with a recursive starting of programs on nodes. Industrial collaborations were carried out with Mandrake, BULL, HP and Microsoft.

5.1.1. KA-Deploy: deployment tool for clusters and grids

KA-DEPLOY is an environment deployment toolkit that provides automated software installation and reconfiguration mechanisms for large clusters and light grids. The main contribution of KA-DEPLOY 2 toolkit is the introduction of a simple idea, aiming to be a new trend in cluster and grid exploitation: letting users concurrently deploy computing environments tailored exactly to their experimental needs on different sets of nodes. To reach this goal KA-DEPLOY must cooperate with batch schedulers, like OAR, and use a parallel launcher like TAKTUK (see below).

5.1.2. Taktuk: parallel launcher

TAKTUK is a tool to launch or deploy efficiently parallel applications on large clusters, and simple grids. Efficiency is obtained thanks to the overlap of all independent steps of the deployment. We have shown that this problem is equivalent to the well known problem of the single message broadcast. The performance gap between the cost of a network communication and of a remote execution call enables us to use a work stealing algorithm to realize a near-optimal schedule of remote execution calls. Currently, a complete rewriting based on a high level language (precisely Perl script language) is under progress. The aim is to provide a light and robust implementation. This development is lead by the MOAIS project-team.

5.1.3. NFSp: parallel file system

When deploying a cluster of PCs there is a lack of tools to give a global view of the available space on the drives. This leads to a suboptimal use of most of this space. To address this problem NFSP was developed, as an extension to NFS that divides file system handling in two components: one responsible for the data stored and the other for the metadata, like inodes, access permission.... They are handled by a server, fully NFS compliant, which will contact associated data servers to access information inside the files. This approach enables a full compatibility, for the client side, with the standard in distributed file systems, NFS, while permitting the use of the space available on the clusters nodes. Moreover efficient use of the bandwidth is done because several data servers can send data to the same client node, which is not possible with a usual NFS server. The prototype has now reached a mature state. Sources are available at http://nfsp.imag.fr.

5.1.4. aIOLi

Modern distributed software uses and creates huge amounts of data with typical parallel I/O access patterns. Several issues, like *out-of-core limitation* or *efficient parallel input/output access* already known in a local context (on SMP nodes for example), have to be handled in a distributed environment such as a cluster.

We have designed AIOLI, an efficient I/O library for parallel access to remote storage in SMP clusters. Its SMP kernel features provide parallel I/O without inter-processes synchronization mechanisms as well as a simple interface based on the classic UNIX system calls (create/open/read/write/close). The AIOLI solution allows us to achieve performance close to the limits of the remote storage system. This was done in several steps:

- Build a local framework that can do aggregation of requests at the application level. This is done by putting a layer between the application and the kernel in charge of delaying individual requests in order to merge them and thus improve performances. The key factor here is to control the delay that should be large enough to discover aggregation patterns but with a limit to avoid excessive waiting times.
- Schedule all I/O requests on a cluster in a global way in order to avoid congestion on a server that leads to bad performances.
- Schedule I/O requests locally on the server so that methods of aggregation and mixing of client requests can be used to improve performances. For that reason AIOLI had to be ported to the kernel and placed at both the VFS level and the lower file system one.

Today, AIOLI compares favorably with the best MPI/IO implementation without any modification of the applications [54] sometimes with a factor of 4. AIOLI can be downloaded from the address http://aioli.imag.fr, both the user library and the Linux kernel module versions.

5.1.5. Gedeon

Gedeon is a middleware for data management on grids. It handles metadata, lists of records made of (attribute, value) pairs, stored in a distributed manner on a grid. Advanced requests can be done on them, using regular expression, and they can be combined in traditional ways, aggregation for example, or used through join operations to federate various sources.

5.1.6. Generic trace and visualization: Paje

This software was formerly developed by members of the Apache project-team. Even if no real research effort is anymore done on this software, many members of the MESCAL project-team use it in their everyday research and promote its use. This software is now mainly maintained by Benhur Stein from Federal University Santa Monica (UFSM), Brazil.

PAJE allows applications programmers to define what is visualized and how new objects should be drawn. To achieve such flexibility, the hierarchy of events and the visualization commands may be defined by the programmers inside the applications. The visualization of parallel execution of ATHA-PAS-CAN applications was achieved without any new addition into PAJE software. Inserting few events trace into the ATHA-PAS-CAN runtime allows the visualization of different facets of the program: application computation time but also user task graph management and scheduling of these tasks. PAJE is also, among others, used to visualize Java program execution and large cluster monitoring. PAJE is actively used by the SIMGRID users' community and the NUMASIS project (see Section 8.2.4).

5.1.7. OAR: a simple and scalable batch scheduler for clusters and grids

OAR is a batch scheduler that emphasizes simplicity, extensibility, modularity, efficiency, robustness and scalability. It is based on a high level conception that reduces drastically its software complexity. Its internal architecture is built on top of two main components: a generic and scalable tool for the administration of the cluster (launch, nodes administration, ...) and a database as the only way to share information between its internal modules. Completely written in Perl, OAR is also extremely modular and straightforward to extend. Thus, it constitutes a privileged platform to develop and evaluate several scheduling algorithms and new kinds of services.

Most known batch schedulers (PBS, LSF, Condor, ...) are of old-fashioned conception, built monolithically, with the purpose of fulfilling most of the exploitation needs. This results in systems of high software complexity (150000 lines of code for OpenPBS), offering a growing number of functions that are, most of the time, not used. In such a context, it becomes hard to control both the robustness and the scalability of the whole system.

The OAR project focuses on robust and highly scalable batch scheduling for clusters and grids. Its main objectives are the validation of grid administration tools such as TAKTUK, the development of new paradigms for grid scheduling and the experimentation of various scheduling algorithms and policies.

The grid development of OAR has already started with the integration of best effort jobs whose purpose is to take advantage of idle times of the resources. Managing such jobs requires a support of the whole system from the highest level (the scheduler has to know which tasks can be canceled) down to the lowest level (the execution layer has to be able to cancel awkward jobs). The OAR architecture is perfectly suited to such developments thanks to its highly modular architecture. Moreover, this development is used for the CiGri grid middleware project.

The OAR system can also be viewed as a platform for the experimentation of new scheduling algorithms. Current developments focus on the integration of theoretical batch scheduling results into the system so that they can be validated experimentally.

5.2. Simulation tools

5.2.1. SimGrid: simulation of distributed applications

SIMGRID implements realistic fluid network models that enable very fast yet precise simulations. SIMGRID enables the simulation of distributed scheduling agents, which has become critical for current scheduling research in large-scale platforms.

Sources and documentations of SIMGRID are available at the following address http://simgrid.gforge.inria.fr/.

5.2.2. ψ and ψ^2: perfect simulation of Markov Chain stationary distribution

ψ and ψ^2 are two software implementing perfect simulation of Markov Chain stationary distributions using the coupling from the past technique. ψ starts from the transition kernel to derive the simulation program while ψ^2 uses a monotone constructive definition of a Markov chain. They are available at http://www-id.imag.fr/Logiciels/psi/.

5.2.3. PEPS

The main objective of PEPS is to facilitate the solution of large discrete event systems, in situations where classical methods fail. PEPS may be applied to the modeling of computer systems, telecommunication systems, road traffic, or manufacturing systems. The software is available at http://www-id.imag.fr/Logiciels/peps/.

5.3. HyperAtlas

The Hyperatlas software has been jointly developed with LSR-IMAG in the framework of the ESPON European project part 3.1 and 3.2. It includes visualization and analysis of socio-economical data in Europe at Nuts 1, Nuts 2 or Nuts 3 level providing analysis of dependence and spatial interaction. This software is available for European partners at http://www-lsr.imag.fr/HyperCarte/.

6. New Results

6.1. Perfect Simulation

Participants: Bruno Gaujal, Brigitte Plateau, Florence Perronnin, Jean-Marc Vincent, Ana Busic.

Perfect simulation enables one to compute samples distributed according to the stationary distribution of the Markov process with no bias. The following sections summarize the various new results obtained using this technique, or on this technique.

6.1.1. Perfect sampling of stationary rewards of Markov chains

In [40], we have illustrated how reward backward coupling improves simulation complexity for the estimation of stationary rewards. Bounds on the coupling time for M/M/1/C have been given and experimental results on a large queueing network validate the practical interest of such an approach.

6.1.2. Perfect simulation and non-monotone Markovian systems

Perfect simulation, or coupling from the past, is an efficient technique for sampling the steady state of monotone discrete time Markov chains. Indeed, one only needs to consider two trajectories corresponding to minimal and maximal state in the system. We show in [44], [26] that even for non-monotone systems one only needs to compute two trajectories: an infimum and supremum envelope. Since the sequence of states obtained by taking infimum (resp. supremum) at each time step does not correspond to a feasible trajectory of the system, envelopes and not feasible trajectories. We show that the envelope approach is efficient for some classes of non-monotone queuing networks, such as networks of queues with batch arrivals, queues with fork and join nodes and/or with negative customers.

Further improvements are going by using the notion of synchronization words in automata theory. It is indeed possible to construct the set of all coupling events of a given Markov chain under the from of an automata. This construction can be done over a state space which is isomorphic to the set of sets of states of the Markov chain, hence in exponential time in general. In some cases however, the complexity can be reduced.

Another direction of research considers a special class of events, namely piecewise space homogeneous events, which cover almost every possible action in queueing networks. In that case, compactification of the space space and linear programming can be used to make computations effective (always) and efficient (in several well identified cases).

6.1.3. Perfect Simulation of Stochastic Automata Networks

The solution of continuous and discrete-time Markovian models is still challenging mainly when we model large complex systems, for example, to obtain performance indexes of parallel and distributed systems. However iterative numerical algorithms, even well-fitted to a multidimensional structured representation of Markov chains, still face the state space explosion problem. Discrete- event simulations can estimate the stationary distribution based on long run trajectories and are also alternative methods to estimate performance indexes of models. Perfect simulation algorithms directly build steady-state samples avoiding the warm-up period and the initial state bias of forward simulations. In [29], we introduce the concepts of backward coupling and the advantages of monotonicity properties and component-wise characteristics to simulate Stochastic Automata Networks (SAN). The main contribution is a novel technique to solve SAN descriptions originally unsolvable by iterative methods due to large state spaces. This method is extremely efficient when the state space is large and the model has dynamic monotonicity because it is possible to contract the reachable state space in a smaller set of maximal states. Component-wise characteristics also contribute to the state space reduction extracting extremal states of the model underlying chain. The efficiency of this technique applied to sample generation using perfect simulation is compared to the overall efficiency of using an iterative numerical method to predict performance indexes of SAN models.

6.2. Tools for Performance Evaluation

Participants: Jean-Michel Fourneau, Brigitte Plateau, Jean-Marc Vincent.

6.2.1. Model Checking

This is collaborative work with Stavros Tripakis (Cadence Research Laboratories)

Exhaustive verification often suffers from the state-explosion problem, where the reachable state space is too large to fit in main memory. For this reason, and because of disk swapping, once the main memory is full very little progress is made, and the process is not scalable. To alleviate this, partial verification methods have been proposed, some based on randomized exploration, mostly in the form of random walks. In [20], we enhance partial, randomized state-space exploration methods with the concept of resource-awareness: the exploration algorithm is made aware of the limits on resources, in particular memory and time. We present a memory-aware algorithm that by design never stores more states than those that fit in main memory. We also propose criteria to compare this algorithm with similar other algorithms. We study properties of such algorithms both theoretically on simple classes of state spaces and experimentally on some preliminary case studies.

6.2.2. Performance characterization of black boxes with self-controlled load injection

This is a collaborative work with Ahmed Harbaoui (France Télécom), Ahmed Harbaoui (France Télécom)

Sizing and capacity planning are key issues that must be addressed by anyone wanting to ensure a distributed system will sustain an expected workload. Solutions typically consist in either benchmarking, or modelling and simulating the target system. However, full-scale benchmarking may be too costly and almost impossible, while the granularity of modelling is often limited by the huge complexity and the lack of information about the system. In [33], we propose a methodology that combines both solutions by first identifying a middle-grain model made of interconnected black boxes, and then to separately characterize the performance and resource consumption of these black boxes. We also propose a component-based supporting architecture, introducing control theory issues in a general approach to autonomic computing infrastructures.

6.2.3. Optical Networks

In [15], we present an approximate analytical method for the evaluation of packet loss probability in synchronous optical packet-switched networks which operate under limited deflection routing with the contention resolution method based on priorities. Packets are lost because they experience too many deflections and stay prohibitively long in the network. The presented results are those for the network in the torus topology of a two-dimensional grid, which operates at a steady state with the uniform load $u \in (0, 1)$.

6.2.4. Stochastic Automata Networks

With excellent cost/performance trade-offs and good scalability, multiprocessor systems are becoming attractive alternatives when high performance, reliability and availability are needed. They are now more popular in universities, research labs and industries. In these communities, life-critical applications requiring high degrees of precision and performance are executed and controlled. Thus, it is important for the developers of such applications to analyze during the design phase how hardware, software and performance related failures affect the quality of service delivered to the users. This analysis can be conducted using modeling techniques such as transition systems. However, the high complexity of such systems (large state space) makes them difficult to analyze.

We have obtained new results on the following aspects:

- the automation of the process to include phase type distribution in SANs [14]
- an efficient algorithm for transient analysis for SANs applied to a GRID 5000 model [24]
- a product form steady-state distribution for SAN with domino synchronizations [32] and for discrete time Markov chains competing over resources [30].
- an algebraic condition for product form in stochastic automata networks without synchronizations [11].
- we analyzed Multiclass G-Networks of Processor Sharing Queues with Resets [31]
- we used an algebraic approach to compute stochastic Bounds for partially generated Markov chains [25].
- the efficient use of Multivalued Decision Diagram (MDD) for SANs. The main difficulty, which was underlined by previous work, is the use of general functions as transition rates or transition probabilities. The term "general" refers to the fact that these functions express dependencies among components. This work is part of the PhD thesis of A. Sales and will be published next year.

6.3. Scheduling

Participants: Jean-Michel Fourneau, Bruno Gaujal, Arnaud Legrand, Jean-François Méhaut.

6.3.1. Minimization of Circuit Registers

This is collaborative work with Jean Mairesse (CNRS LIAFA)

In [12], we address the following problem: given a synchronous digital circuit, is it possible to construct a new circuit computing the same function as the original one but using a minimal number of registers? The construction of such a circuit can be done in polynomial time and is based on a result of Orlin for one periodic bi-infinite graphs showing that the cardinality maximum flow is equal to the size of a minimum cut. The idea is to view such a graph as the unfolding of the dependences in a digital circuit.

6.3.2. Optimal end-to-end routing for networks with multiplexing

In [10] we show how Network Calculus can be used to compute the optimal route for a flow (w.r.t. end-to-end guarantees on the delay or the backlog) in a network in the presence of cross-traffic. When cross-traffic is independent, the computation is shown to boil down to a functional shortest path problem. Under usual assumptions (concave arrival curves and convex service curves), a simple min-max lemma enables to solve the problem in polynomial time by a sequence of classical shortest path computations. When cross-traffic perturbs the main flow over more than one node and under blind multiplexing, one can take into account the “Pay Multiplexing Only Once” (PMOO) phenomenon. It enables to improve bounds on delays and backlogs, but it makes the computation more involved. We provide a formula which gives a service curve for a path with PMOO conditions. It introduces a new multi-dimensional Network Calculus operator and generalizes previous formula on PMOO. Moreover when arrival (resp. service) curves are affine (resp. convex), we describe an efficient algorithm to compute this formula. We finally show how to adapt our routing algorithms to have optimal end-to-end guarantees for these new bounds on delays and backlogs which take into account PMOO.

6.3.3. Analyzing Weighted Round Robin policies with a Stochastic Comparison Approach

In [8], we study queuing delays for Weighted Round Robin (WRR) scheduling policies. The delay characteristics of these policies can be estimated through worst-case bounds as other fair queuing scheduling policies. However, these deterministic bounds are in general not accurate and they require that the input process has been shaped. Under Markovian arrival hypothesis, the delay of WRR policies can be evaluated through Markovian numerical analysis. Nevertheless, this analysis is limited to small parameter sizes because of the state space explosion problem. We propose to apply the stochastic comparison approach. We build a bounding model based on the aggregation of sessions, and we show that under the same arrivals, the packet delays in the bounding model are larger in the $\leq_{st}$ stochastic order sense than the packet delays in the original model. Due to this aggregation, the state space size is drastically reduced and we are able to provide stochastic delay bounds for WRR policies. We discuss the accuracy of the proposed bounds through numerical examples.

6.3.4. Scheduling Deadline Constrained Checkpointing on Virtual Clusters

In [47], [42], we consider a context where the available resources of the Intranet of a company are used as a virtual cluster for scientific computation, during the idle periods (nights, weekends, holidays,...). Generally, these idle periods do not permit to carry out completely the computations. For instance, a workstation mobilized during the night must be released in the morning to make it available for the employee, even if the application running on it is not completed. It is therefore necessary to save the context of uncompleted applications for possible restart. Hereafter, we assume that the computations running on the workstations are independent from each other. The checkpointing mechanism which ensures the continuity of applications is subject to resource constraints : the network bandwidth, the disk bandwidth and the delay T imposed for releasing the workstations. We first show that the designing of a scheduling strategy which optimizes resource consumption while taking into account the above constraints, can be formalized as a variant of the classical 0/1 knapsack problem. Then, we propose an algorithm whose implementation does not have a significant overhead on checkpointing mechanisms. Experiments carried out on a real cluster show that this algorithm performs better than the naive scheduling algorithm which selects the applications one after the other in order of decreasing amount of resource consumption.

6.4. Middleware and Experimental Testbeds

Participants: Olivier Richard, Yves Denneulin, Jean-François Méhaut.

6.4.1. Parallel Implementation of the STL for multi-core machines

In [46], we have proposed the PaSTeL library, a parallel implementation of a subset of the STL, the standard C++ library. PaSTeL is based on a new model of parallel programming and on a pragmatic work-stealing execution model. In particular, we use optimized thread synchronization and activation mechanisms. The performances of PaSTeL have been evaluated on both a standard bicore workstation and a 16-core parallel machine and are far better than other existing parallel implementation of the STL, even on small datasets.

6.4.2. Lightweight Emulation to Study Peer-to-Peer Systems

The current methods used to test and study peer-to-peer systems (namely modeling, simulation, or execution on real testbeds) often show limits regarding scalability, realism and accuracy. In [18], [6] we present the design and the evaluation of P2PLab, a framework to study peer-to-peer systems by combining emulation (use of the real studied application within a configured synthetic environment) and virtualization. P2PLab is scalable (it uses a distributed network model) and has good virtualization characteristics (many virtual nodes can be executed on the same physical node by using process level virtualization). Experiments with the BitTorrent file sharing system complete this work and demonstrate the usefulness of this platform.

6.4.3. Experiment Engine for Lightweight Grids

In [38], we present a case study conducted on the Grid'5000 platform, a lightweight grid. The goal was to make a rather simple experiment, and study how difficult it was to carry out correctly, i.e. to be correct, reproducible and efficient. The work shows that despite the precautions taken, many parameters that could have an effect on the result were at first overlooked. It also shows that benchmarking plays a key role on making an experiment correct and reproducible. The process is in the end extremely tedious, and stresses the need for new tools to help users. We have thus presented a methodology to get correct results on grid architecture, to identify relevant problems and to propose an infrastructure that answers part of the problems encountered during experiments. Additionally, pieces of this infrastructure have been built.

6.5. Distributed Computing Platforms: Measurements and Models

Participants: Yves Denneulin, Derrick Kondo, Jean-François Méhaut, Olivier Richard, Jean-Marc Vincent.

6.5.1. Nanosimulation

In [45], we have analyzed an electronic structure simulation application. The simulation of the structure and of the material property and molecules is based on quantum mechanics and more specifically on Shrodinger's equation. Our aim was to characterize as accurately as possible the performance and the behavior of this application so as to determine its optimal platform configuration. The cluster nodes are hierarchical multi-core SMPs and share a hierarchical memory (NUMA). These experiments have been conducted on two types of NUMA SMPs based either on Intel Itanium or on AMD Opteron CPUs using three different Fortran compilers (two commercial ones and a free one).

6.5.2. Impact of Data Sharing in Load Balancing

In [43], we have evaluated data replication and application scheduling strategies on a grid. To this end we have developed a prototype implementing a data distribution based either on a central server or on BitTorrent. This prototype enabled us to measure the impact of data sharing on the performance of scheduling strategies.

6.5.3. Memory Affinity

In [37], we have started modeling and analyzing memory affinity on NUMA architectures.

6.5.4. Resources Availability for Peer-to-Peer Systems

In [35], we measure and characterize the time dynamics of availability in a large-scale Internet-distributed system with over 110,000 hosts from SETI@home. Our characterization focuses on identifying patterns of correlated availability. We determine scalable and accurate clustering techniques and distance metrics for automatically detecting significant availability patterns. By means of clustering, we identify groups of resources with correlated availability that exhibit similar time effects. Then we show how these correlated clusters of resources can be used to improve resource management for parallel applications in the context of volunteer computing.

In [21], we create predictive models of this new data set. Increasingly services are being deployed over large-scale computational and storage infrastructures. To meet ever-increasing computational demands and to reduce both hardware and system administration costs, these infrastructures have begun to include Internet resources distributed over enterprise and residential broadband networks. As these infrastructures increase in scale to hundreds of thousands to millions of resources, issues of resource availability and service reliability inevitably emerge. We determine and evaluate predictive methods that ensure the availability of a collection of resources. With the aforementioned data set, we show how to reliably and efficiently predict that a collection of N hosts will be available for T time. The results indicate that by using replication it is feasible to deploy enterprise services or applications even on such volatile resource pools.

6.5.5. Predictive Models for Bandwidth Sharing in High Performance Clusters

Using MPI as the communication interface, one or several applications may introduce complex communication behaviors over the network cluster. This effect is increased when nodes of the cluster are multi-processors, and where communications can come or go from the same node with a common interval time. Our goal is to understand those behaviors and to build a class of predictive models of bandwidth sharing, knowing, on the one hand the flow control mechanisms and, on the other hand, a set of experimental results. In [17], [39], we present experiences that show how bandwidth sharing on Gigabit Ethernet, Myrinet 2000 and Infiniband networks before to introduce the models for Gigabit Ethernet and Myrinet 2000 networks.

6.6. Multi-User Systems

Participants: Bruno Gaujal, Arnaud Legrand, Corinne Touati, Jean-Marc Vincent.

6.6.1. Multi-Agent Systems

We have started working on the evaluation of Multi-Agent Systems (MAS) at the level of their interaction. Two problems that may be a bias in the evaluation and measurement of interaction are discussed in [34]. The first one is the difference between the quantities of information carried by a unit of interaction in two systems having different architectures. The second one concerns the interaction units that are received and cannot be exploited by the agent. In this work, an evaluation based on the weight of the information brought by an interaction is suggested. In order to achieve this, a MAS model, on which the evaluation is based, is defined. Then, the different problems and solutions which will help to evaluate the interaction are studied. Finally, the approach is applied on two different implementations that solve the same problem

6.6.2. User-Network Association in Multi-Technology Wireless Networks

Recent mobile equipment (as well as the norm IEEE 802.21) now offers the possibility for users to switch from one technology to another (vertical handover). This allows flexibility in resource assignments and, consequently, increases the potential throughput allocated to each user. In [51], [28], we design a fully distributed algorithm based on trial and error mechanisms that exploits the benefits of vertical handover by finding fair and efficient assignment schemes. On the one hand, mobiles gradually update the fraction of data packets they send to each network based on the rewards they receive from the stations. On the other hand, network stations send rewards to each mobile that represent the impact each mobile has on the cell throughput. This reward function is closely related to the concept of marginal cost in the pricing literature. Both the station and the mobile algorithms are simple enough to be implemented in current standard equipment. Based on tools from evolutionary games, potential games and replicator dynamics, we analytically show the convergence of the algorithm to solutions that are efficient and fair in terms of throughput. Moreover, we show that after convergence, each user is connected to a single network cell which avoids costly repeated vertical handovers. Several simple heuristics based on this algorithm are proposed to achieve fast convergence. Indeed, for implementation purposes, the number of iterations should remain in the order of a few tens. We also compare, for different loads, the quality of their solutions.

6.6.3. Decentralized Scheduling Algorithm for Multiple Bag-of-tasks Application Scheduling on Grids

We have designed a fully decentralized algorithm for fair resource sharing between multiple bag-of-tasks applications in a grid environment [49], [23]. This algorithm is inspired from related work on multi-path routing in communication network. An interesting feature of this algorithm is that it allows the choice of wide variety of fairness criteria and achieves both optimal path selection and flow control. In addition, this algorithm only requires local information at each slave computing tasks and at each buffer of the network links while minimal computation is done by the schedulers. A naive adaptation is unstable and inefficient though. Fortunately, a simple and effective scaling mechanism is sufficient to circumvent this issue. This scaling mechanism is motivated by a careful study of the subtle differences with the classical multi-path routing problem. We have shown its efficiency through a detailed analysis of a simple simulation but we are still working on a more wide and in-depth analysis.

6.7. On-demand Geographical Maps

Participant: Jean-Marc Vincent.

The new results regarding on-demand geographical maps are twofold.

- The potential methods have been developed in the HyperSmooth software and applied in the European ESPON project [19].
- The HyperSmooth software architecture has been presented in China [36].

6.8. Discrete Structures

Participants: Yves Denneulin, Bruno Gaujal.

6.8.1. Distributing Labels on Infinite Trees

Sturmian words are infinite binary words with many equivalent definitions: They have a minimal factor complexity among all aperiodic sequences; they are balanced sequences (the labels 0 and 1 are as evenly distributed as possible) and they can be constructed using a mechanical definition. All this properties make them good candidates for being extremal points in scheduling problems over two processors. In [52], we consider the problem of generalizing Sturmian words to trees. The problem is to evenly distribute labels 0 and 1 over infinite trees. We have shown that (strongly) balanced trees exist and can also be constructed using a mechanical process as long as the tree is irrational. Such trees also have a minimal factor complexity. Therefore they bring the hope that extremal scheduling properties of Sturmian words can be extended to such trees, as least partially. Such possible extensions are illustrated by one such example in scheduling theory.

6.8.2. Spanning Trees across Hypercubes

This is collaborative work with Maurice Tchuente (Université de Yaoundé), Paulin Yonta (Université de Yaoundé), Jean-Michel Nlong II (Université de Ngaoundéré)

Given an undirected and connected graph G, with a non-negative weight on each edge, the Minimum Average Distance (MAD) spanning tree problem is to find a spanning tree of G which minimizes the average distance between pairs of vertices. This network design problem is known to be NP-hard even when the edge-weights are equal. In [16], we make a step towards the proof of a conjecture stated by A.A. Dobrynin, R. Entringer and I. Gutman in 2001, and which says that the binomial tree B_n is a MAD spanning tree of the hypercube H_n . More precisely, we show that the binomial tree B_n is a local optimum with respect to the 1-move heuristic. We also present a greedy algorithm which produces good solutions for the MAD spanning tree problem on regular graphs such as the hypercube and the torus.

7. Contracts and Grants with Industry

7.1. CIFRE with BULL, 06-09

Yiannis Georgiou is doing his PhD thesis in a CIFRE contract with the BULL company. His work started in September 2006, and he will finish in September 2009. The focus of his research is batch scheduling on Grids.

7.2. CIFRE with France Télécom R&D, 06-09

Ahmed Harbaoui is doing his PhD thesis in a CIFRE contract with the France Télécom R&D company. His work started in September 2006, and he will finish in September 2009. He is interested in load injection and performance evaluation issues in networks.

7.3. CIFRE with STMicroelectronics, 06-10

Carlos Rojas is doing his PhD thesis under a CIFRE contract with STMicroelectronics. He started in September 2007 and will finish in September 2010. The objective of his thesis is to develop methods and tools for multiprocessor embedded applications.

7.4. Sceptre with STMicroelectronics, (Divisions STS and HEG), INRIA Rhône-Alpes (MOAIS, Mescal, Arenaire, CompSys), TIMA/SLS, Verimag, CAPS-Entreprise and IRISA (CAPS) 06-10

Sceptre is a minalogic project, supported by the Pole de Competitivite Minalogic. Global competitiveness cluster Minalogic fosters research-led innovation in intelligent miniaturized products and solutions for industry. Located in Grenoble, France, the cluster channels in a single physical location a range of highly-specialized skills and resources from knowledge creation to the development and production of intelligent miniaturized services for industry. Sceptre main objective is to provide SoC implementation techniques, using novel approaches originating from both multiprocessor programming and reconfigurable processors. The application domain is distributed multimedia code optimization.

Our work is focused on tools and methods to develop embedded systems. The main working directions are software and hardware integration, scalable and configurable architectures, real time constraints, heterogeneous multiprocessing, and load-balancing.

7.5. Real-Time-At -Work

RealTimeAtWork.com is a startup from INRIA Lorraine created in December 2007. Some members of Mescal are scientific partners in the startup. Its main target is to provide software tools for solving real time constraints in embedded systems, particularly for superposition of periodic flows. Such flows are typical in automotive and avionics industries who are the privileged potential users of the technologies developed by RealTimeAtWork.com

7.6. CILOE with BULL, Compagnie des Signaux, TIMA, CEA-LETI, LIG, Edxact, Infiniscale, Probayes, SCelectronique, 06-10

The increasingly miniaturization of components and the ever-increasing complexity of electronic circuits for communication systems requires a set of sophisticated tools for design and simulation. These tools in turn often require immense computational resources, sometimes more than several orders of magnitude above the performance of a desktop PC or a workstation. These tools are so compute-intensive that they require supercomputers, clusters and grids. However, these types of computing resources are often not within the reach of PME's (relatively small companies or startups) in the semiconductor industry and sometimes even large companies, not only because of the cost of infrastructure, but also because of the lack of adequate methods and technologies for high performance computing.

In the association of Minalogic, there are about twenty PME's that develop CAD software, and other companies in the field of embedded systems, the design of electronic circuits, and the simulation process. The most advanced companies utilize high performance computing, and the others will have to do so in 2 or 3 years. All of these companies are confronted with a notable lack of services and facilities for intensive computing, which heavily affect their competitiveness and speed of development.

It is in this context that the partners of this CILOE project propose to design and develop a complete computational infrastructure, including methodologies, software, and security mechanisms. This infrastructure will contribute decisively to the development and visibility of the international PME partners in the project. It will be an essential tool for a sustainable boost in the sector of electronic CAD, embedded software and high-performance simulation and moreover, facilitate growth for all companies in the electronics industry in Alpes region.

This project has three main objectives that will allow industry to leverage large-scale compute-intensive platforms:

- Reduce the delay in the development of reliable software of the industry partners (Jivaro for Edxact, ProBayes-BT for Probayes, Stressio for SC Online). The validation of software improvements requires numerous test cases of modest size but also test cases of much larger size. For example, the biggest test case (15 GB approximately) for the software Jivaro of the company Edxact requires computation on the order of days. Often, the long duration of these computations can delay the validation of software. The goal here is to improve the competitiveness of local companies so that they can provide more quickly new versions of their software that has been completely validated in a number of tests.
- Develop highly parallel versions of software of the PME/PMI partners. The targeted architectures here are clusters of multi-core machines and specialized processors (system-on-a-chip multiprocessors, NoC-; Cell). This technological gain for business partners (Edxact, ProBayes) will enhance their competitiveness.
- Experiment with services for enabling resource access by applications. This would be based on the principles of IaaS (Infrastructure as a Service) and SaaS (Software as a Service). In the models of IaaS and SaaS, customers of the PME partners do not have to pay for the construction and maintenance of the entire infrastructure and software licenses. Instead, the customers only pay for their direct use. Once the infrastructure and services are deployed, customer access is enabled through a simple Web interface, which will allow PME's to cheaply target a global market.

8. Other Grants and Activities

8.1. Regional initiatives

8.1.1. CIMENT

The CIMENT project (Intensive Computing, Numerical Modeling and Technical Experiments, http://ciment.ujf-grenoble.fr/) gathers a wide scientific community involved in numerical modeling and computing (from numerical physics and chemistry to astrophysics, mechanics, bio-modeling and imaging) and the distributed computer science teams from Grenoble. Several heterogeneous distributed computing platforms were set up (from PC clusters to IBM SP or alpha workstations) each being originally dedicated to a scientific domain. More than 600 processors are available for scientific computation. The MESCAL project-team provides expert skills in high performance computing infrastructures.

8.1.2. Grappe200 project

MENRT-UJF-INPG, RhÃ´ne-Alpes Region, INRIA , ENS-Lyon have funded a cluster composed of 110 bi-processors Itanium2 connected with a Myrinet (donation of MyriCom) high performance network. This project is lead by MESCAL, MOAIS, GRAAL and SARDES. It is part of the CIMENT project which aims at building high performance distributed grids between several research labs (see above).

8.1.3. Cluster Région

The MESCAL project-team is a member of the regional "cluster" project on computer science and applied mathematics, the focus of its participation is on handling large amount of data large scale architecture. Other members of this subproject are the INRIA GRAAL project-team, the LSR-IMAG and IN2P3-LAPP laboratories.

8.2. National initiatives

8.2.1. Aladdin-G5K, 2008-2011, ADT

Partners: INRIA FUTURS, INRIA Sophia, IRISA, LORIA, IRIT, LABRI, LIP, LIFL.

After the success of the Grid'5000 project of the ACI Grid initiative led by the French ministry of research, INRIA is launching the ALADDIN project to further develop the Grid'5000 infrastructure and foster scientific research using the infrastructure.

ALADDIN will build on Grid'5000's experience to provide an infrastructure enabling computer scientists to conduct experiments on large scale computing and produce scientific results that can be reproduced by others. ALADDIN focus on the following challenges :

1. Transparent, safe and efficient large scale system utilization and programming
2. Providing service agreement to users in large scale parallel and distributed systems
3. Providing confidence to the user about the infrastructure
4. Efficient exploitation of highly heterogeneous and hierarchical large-scale systems
5. Efficient and scalable composition and orchestration of services
6. Modeling of large scale systems and validation of their simulators
7. Scalable applications for large scale systems
8. Dynamic interconnection of autonomous and heterogeneous resources
9. Efficiently manage very large volumes of information (search, mining, classification, secure storage and access, etc) for a wide spectrum of applications areas (web applications, image processing, health, environment, etc).

Mescal members are particularly involved in topics 1, 3, 4, and 6.

8.2.2. POPEYE, 2008-2009, ARC

Partners: INRIA Maestro, INRIA TOSCA, INRA, UMPC, LIA, Polytech Nice Sophia-Antipolis.

The MESCAL participates in the Popeye INRIA ARC, lead by Eltan Altman of the INRIA Maestro project-team. The project focuses on the behavior of large complex systems that involve interactions among one or more populations. By population we mean a large set of individuals, that may be modeled as individual agents, but that we will often model as consisting of a continuum of non-atomic agents. The project brings together researchers from different disciplines: computer science and network engineering, applied mathematics, economics and biology. This interdisciplinary collaborative research aims at developing new theoretical tools as well as at their applications to dynamic and spatial aspects of populations that arise in various disciplines, with a particular focus on biology and networking.

8.2.3. DSLLab, 2005-2008, ANR Jeunes Chercheurs

Partners: INRIA-FUTURS.

DSLlab is a research project aiming at building and using an experimental platform about distributed systems running on DSL Internet. The objective is twofold:

- provide accurate and customized measures of availability, activity and performances in order to characterize and tune the models of the ASDL resources;
- provide a validation and experimental tool for new protocols, services and simulators and emulators for these systems.

DSLlab consists of a set of low power, low noise computers spread over the ASDL. These computers are used simultaneously as active probes to capture the behavior traces, and as operational nodes to launch experiments. We expect from this experiment a better knowledge of the behavior of the ASDL and the design of accurate models for emulation and simulation of these systems, which represents now a significant capability in terms of storage and computing power.

8.2.4. NUMASIS, 2005-2008, ANR Calcul Intensif et Grilles de Calcul

Future generations of multiprocessors machines will rely on a NUMA architecture featuring multiple memory levels as well as nested computing units (multi-core chips, multi-threaded processors, multi-modules NUMA, etc.). To achieve most of the hardware's performance, parallel applications need powerful software to carefully distribute processes and data so as to limit non-local memory accesses. The ANR NUMASIS[2] project aims at evaluating the functionalities provided by current operating systems and middleware in order to point out their limitations. It also aims at designing new methods and mechanisms for an efficient scheduling of processes and a clever data distribution on such platforms. These mechanisms will be implemented within operating systems and middleware. The target application domain is seismology, which is very representative of the needs of computer-intensive scientific applications.

8.2.5. ALPAGE, 2005-2008, ANR Masses de Données

The new algorithmic challenges associated with large-scale platforms have been approached from two different directions. On the one hand, the parallel algorithms community has largely concentrated on the problems associated with heterogeneity and large amounts of data. Algorithms have been based on a centralized single-node, responsible for calculating the optimal solution; this approach induces significant computing times on the organizing node, and requires centralizing all the information about the platform. Therefore, these solutions clearly suffer from scalability and fault tolerance problems.

On the other hand, the distributed systems community has focused on scalability and fault-tolerance issues. The success of file sharing applications demonstrates the capacity of the resulting algorithms to manage huge volumes of data and users on large unstable platforms. Algorithms developed within this context are completely distributed and based on peer-to-peer communications. They are well adapted to very irregular applications, for which the communication pattern is unpredictable. But in the case of more regular applications, they lead to a significant waste of resources.

The goal of the ALPAGE project is to establish a link between these directions, by gathering researchers (Mescal, LIP, LORIA, LaBRI, LIX, LRI) from the distributed systems and parallel algorithms communities. More precisely, the objective is to develop efficient and robust algorithms for some elementary applications, such as broadcast and multicast, distribution of tasks that may or may not share files, resource discovery. These fundamental applications correspond well to the spectrum of the applications that can be considered on large scale, distributed platforms.

8.2.6. SMS, 2005-2008, ANR

The ACI SMS, "Simulation et Monotonie Stochastique en évaluation de performances", is composed by two teams: Performance Evaluation team from PRiSM Laboratory (ACI Leader) and the MESCAL project-team. The main objective is to study monotonicity properties of computer systems models in order to speed up the simulations and estimate performance indexes more accurately.

The composition formalisms we have contributed to develop during the recent years allow to build large Markov chains associated to complex systems in order to analyze their performance. However, it is often impossible to solve the stationary or transient distributions. Analytical methods and simulations fail for different reasons.

However brute performances are not really useful. We need the proof that the system is better than an objective. Therefore it is natural to use comparison of random variables and sample-paths. Two important concepts appear: stochastic ordering and stochastic monotony. We chose to develop these two important concepts and apply them to perfect simulation, distributed simulation and product form queuing network. These concepts seem to appear frequently in various techniques in performance evaluation. Using the monotony property, one can reduce the computation time for perfect simulation with coupling from the past. Coupling from the past allows to sample the steady-state distribution in a finite time. Thus we do not encounter the same stopping

[2]NUMASIS: Adapting and Optimizing Applicative Performance on NUMA Architectures: Design and Implementation with Applications in Seismology

problem that holds for ordinary simulations. Furthermore, some results show that the monotony property is often present in queuing network even if they do not have product form. We simply have to renormalize them to let the property appear. Using both properties, it is also possible to derive distributed simulations which will be more efficient. We will develop two ideas: sample-path transformations to avoid rollback in optimistic simulations (and we compute a bound) and regenerative simulations.

Finally, these concepts can be used for product form queuing network to explain why some transformation applied on customer synchronization can provide product form solution, and also how we can compute a solution of the traffic equation when they are unstable.

8.2.7. DOCCA, 2007-2011 ANR Jeunes Chercheurs

The race towards the design and development of scalable distributed systems offers new opportunities to applications, in particular as far as scientific computing, databases, and file sharing are concerned. Recently many advances have been done in the area of large-scale file-sharing systems, building upon the peer-to-peer paradigm that somehow seamlessly responds to the dynamicity and resilience issues. However, achieving a fair resource sharing amongst a large number of users in a distributed way is clearly still an open and active research field. For all previous issues there is a clear gap between

- widely deployed systems as peer-to-peer file-sharing systems (KaZaA, Gnutella, EDonkey) that are generally not very efficient and do not propose generic solutions that can be extended to other kind of usage;
- academic work with generally smart solutions (probabilistic routing in random graphs, set of node-disjoint trees, lagrangian optimization) that sometimes lack a real application.

Up until now, the main achievements based on the peer-to-peer paradigm mainly concern file- sharing issues. We believe that a large class of scientific computations could also take advantage of this kind of organization. Thus our goal is to design a peer-to-peer computing infrastructure with a particular emphasis on the fairness issues. In particular, the objectives of the ANR DOCCA[3] project are the following:

- to combine theoretical tools and metrics from the parallel computing community and from the network community, and to explore algorithmic and analytical solutions to the specific resource management problems of such systems.
- to design a P2P architecture based on the algorithms designed in the second step, and to create a novel P2P collaborative computing system.

We expect the following results from this project:

- to provide user synthetic models to the scientific community that can be used as an input in modeling, simulation and experimentation of P2P collaborative computing systems.
- to provide optimal strategies and resource management algorithms in P2P collaborative computing.
- to design a decentralized protocol that implements the optimal strategies for the target user models.
- to implement a prototype and validate the approach on an experimental platform.

8.2.8. Check-bound, 2007-2009 ANR SETIN

Partners: University of Paris I.

The increasing use of computerized systems in all aspects of our lives gives an increasing importance on the need for them to function correctly. The presence of such systems in safety-critical applications, coupled with their increasing complexity, makes indispensable their verification to see if they behaves as required . Thus the model checking which is the automated manner of formal verification techniques is of particular interest. Since verification techniques have become more efficient and more prevalent, it is natural to extend the range of models and specification formalisms to which model checking can be applied. Indeed the behavior of

[3] Design and Optimization of Collaborative Computing Architectures

many real-life processes is inherently stochastic, thus the formalism has been extended to probabilistic model checking. Therefore, different formalisms in which the underlying system has been modeled by Markovian models have been proposed.

Stochastic model checking can be performed by numerical or statistical methods. In model checking formalism, models are checked to see if the considered measures are guaranteed or not. We apply Stochastic Comparison technique for numerical stochastic model checking. The main advantage of this approach is the possibility to derive transient and steady-state bounding distributions as well as the possibility to avoid the state-space explosion problem. For the statistical model checking we study the application of perfect simulation by coupling in the past. This method has been shown to be efficient when the underlying system is monotonous for the exact steady-state distribution sampling. We consider to extend this approach for transient analysis and to model checking by means of bounding models and the stochastic monotonicity. As one of the most difficult problems for the model checking formalism, we also study the case when the state space is infinite. In some cases, it would be possible to consider bounding models defined in finite state space.

8.2.9. ACI blanche MEG 2007-2010

The "ACI blanche" MEG, is composed of two teams: physicists working on electromagnetism from the LAAS (Toulouse) and the MESCAL project-team. The main objective is to study scaling properties in electromagnetism simulation applications and grids. The first results are promising. They demonstrate that the tools developed by Mescal on large data storage and middleware for deployment on clusters and grids are appropriate for that kind of application.

8.3. International Initiatives

8.3.1. Europe

ESPON : The MESCAL project-team participates to the ESPON (European Spatial Planning Observation Network) http://www.espon.lu/ It is involved in the action 3.1 on tools for analysis of socio-economical data. This work is done in the consortium hypercarte including the laboratories LSR-IMAG (UMR 5526), Géographie-cité (UMR 8504) and RIATE (UMS 2414). The Hyperatlas tools have been applied to the European context in order to study spatial deviation indexes on demographic and sociological data at nuts 3 level.

8.3.2. Africa

Cameroon : MESCAL takes part in the SARIMA[4] project and more precisely with the University of Yaoundé 1. Two Cameroon students (Jean-Michel NLong 2 and Blaise Yenké) are preparing their PhD in cotutelle (joint and remote supervision) with Professor Maurice Tchuenté. SARIMA also funded Adamou Hamza to prepare his Master Thesis during three months in the MESCAL project-team. SARIMA proposed J-F Méhaut to give a course on Operating System and Networks at Master Research Students.

8.3.3. South America

- DIODE (2006-2008) Associate Team funded by INRIA with the MOAIS project-team of INRIA, and the Brazilian University UFRGS. The goal of this project is to design and develop programming tools for grid and clusters for virtual reality. This collaboration was initiated 10 years ago, and has greatly affected the activities (doctoral, publications and joint production software) of the Apache project-team, from which MOAIS and MESCAL were formed. In particular, four PhD Brazilian students have joined the MESCAL project-team as a result of this long-standing collaboration.
- CAPES/COFECUB grant (2006-2008) with the UFRGS, Porto Alegre, Brazil around grid and PC clusters.

[4] Soutien aux Activités de Recherche Informatique et Mathématiques en Afrique http://www-direction.inria.fr/international/AFRIQUE/sarima.html

- ECOS grant (2007-2009) Colombia: joint project with the universities of Los Andes, Bogota, and UIS, Bucaramanga, on the topic of grids for computation and data management.

8.3.4. Pacific and South Asia

Corinne Touati is the INPG correspondent for student exchanges with Japan and has visited many Japanese universities to ease these exchanges.

8.4. High Performance Computing Center

8.4.1. The ICluster2, the IDPot and the new Digitalis Platforms

The MESCAL project-team manages a cluster computing center on the Grenoble campus. The center manages different architectures: a 48 bi-processors PC (ID-POT), and the center is involved with a cluster based on 110 bi-processors Itanium2 (ICluster-2) and another based on 34 bi-processor quad-core XEON (Digitalis) located at INRIA. The three of them are integrated in the Grid'5000 grid platform.

More than 60 research projects in France have used the architectures, especially the 204 processors Icluster-2. Half of them have run typical numerical applications on this machine, the remainder has worked on middleware and new technology for cluster and grid computing. The Icluster-2 has been stopped this year as it was getting obsolete and has been replaced by the Digitalis platform. The Digitalis cluster is also meant to replace the Grimage platform in which the MOAIS project-team is very involved.

8.4.2. The BULL Machine

In the context of our collaboration with BULL (LIPS, NUMASIS), the MESCAL project-team acquired a Novascale NUMA machine. The configuration is based on 8 Itanium II processors at 1.5 Ghz and 16 GB of RAM. This platform is mainly used by the BULL PhD students. This machine is also connected to the CIMENT Grid.

8.4.3. GRID 5000 and CIMENT

The MESCAL project-team is involved in development and management of Grid'5000 platform. The Digitalis and IDPot clusters are integrated in Grid'5000. Moreover, these two clusters take part in CIMENT Grid. More precisely, their unused resources may be exploited to execute jobs from partners of CIMENT project (see Section 8.1.1).

9. Dissemination

9.1. Leadership within the scientific community

9.1.1. Tutorials

Researchers of the MESCAL project-team have been invited to give tutorials on critical research subjects in international conferences:

- Jean-Marc Vincent gave a tutorial on perfect sampling with applications to queueing networks at the QEST conference [41].
- Arnaud Legrand gave a tutorial with Martin Quinson on simulation of large-scale distributed systems at the CCGrid conference.

9.1.2. Conference and Workshop Organization

Researchers of the MESCAL project-team organized and hosted several conferences and workshops:

- Ninth Workshop on Performance Evaluation (AEP9) (Jean-Marc Vincent, Chair)
- Game Theory for Analysis and Optimization of Computer Systems (GameComp) (Corinne Touati, Chair)
- Fourth International BOINC Workshop on Volunteer Computing and Distributed Thinking (Derrick Kondo, Co-chair)

9.1.3. Conference and Workshop Chairing

Researchers of the MESCAL project-team have been chairs of the following conferences or workshops:

- Workshop on Desktop Grids and Volunteer Computing (Derrick Kondo, program chair).

9.1.4. Program committees

Researchers of the MESCAL project-team have been program committee members of the following conferences or workshops:

- Rencontres Francophones du Parallélisme 2008, RenPar'18, Fribourg.
- 2nd Workshop on System-Level Virtualisation for High Performance Computing (HPCVirt 2008), Glasgow.
- EuroPVM/MPI 2008, Dublin.
- IEEE/ACM International Conference on Grid Computing (Grid), Tsukuba.
- Workshop on Desktop Grids and Volunteer Computing, Miami.
- International Workshop on Global and Peer-to-Peer Computing, Lyon.
- IEEE Wireless Communications and Networking Conference, Las Vegas.
- Third International Conference on Performance Evaluation Methodologies and Tools (ValueTools), Athens.
- Second Workshop on Network Control and Optimization, Paris.

9.1.5. Thesis defense

- Lucas Nussbaum defended his PhD on December 4th, 2008 in Grenoble. Thesis committee: Didier Donsez, Franck Cappello, Pierre Sens, Pascale Vicat-Blanc Primet, Jean-Francois Méhaut, Olivier Richard [6]
- Jean-Michel N'Long 2 defended his PhD on July 2008 in Yaoundé 1 (Cameroon). Thesis committee: Brigitte Plateau, Jean-Francois Méhaut, [5].

9.1.6. Thesis committees

Researchers of the MESCAL project-team have served on the following thesis committees:

- Arnaud Legrand served on the thesis committee of Matthieu Pérottin (University of Tours).
- Bruno Gaujal served on the thesis committee of Jean-Vivien Millo (University of Nice and INRIA Sophia) as a reviewer.
- Corinne Touati served on the thesis committee of Dinesh Kumar (University of Nice and INRIA Sophia) and Jean-Marc Kelif (Orange Labs / Telecom ParisTech).
- Jean-FranÃ§ois Méhaut served on the thesis committee of Jean-Michel Nlong (University of Yaoundé), Loic Strus (University of Grenoble), Jean-Baptiste Ernst-Desmuler (University of Montbelliard), Elisabeth Brunet (University of Bordeaux), and Salam Traboulsi (University of Toulouse).

9.1.7. Members of editorial board

9.1.8. Grenoble's Seminar on performance evaluation

This seminar is organized by Jean-Marc Vincent and Bruno Gaujal. It is tightly coupled with the PAGE group and its main goal is to organize meetings between the various researchers of Grenoble using the same kind of mathematical tools (stochastic models, queuing networks, Petri networks, stochastic automata, Markovian process and chains, (max,+) algebra, fluid systems, ...). On the long term, this seminar should lead to inter-laboratory working groups on precise themes. More information is available at

9.2. Teaching

Members of the MESCAL team are actively involved in teaching. Their activities are balanced between graduate students and post-graduate students. Here are a few examples of their responsibilities:

- **2nd year of Research Master of Paris (MPRI)** Bruno Gaujal gives a course on discrete event dynamic systems.
- **2nd year of International Research Master of Grenoble (MOSIG)** Here is a list of courses taught by researchers of the MESCAL project-team:
  - Cluster architectures for high-performance computing and high throughput data management.
  - Data measurement and analysis for network and operating systems performance evaluation.
  - Modeling and simulation for network and operating systems performance evaluation.
  - Building parallel and distributed applications (contributor).
  - Algorithms and basic techniques for parallel computing (contributor).
- **2nd year of Research Master (Yaoundé)** Operating systems and networks.
- **MagistÃ¨re d'informatique Licence (Université Joseph Fourier)**

10. Bibliography

Major publications by the team in recent years

[1] E. ALTMAN, B. GAUJAL, A. HORDIJK. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity*, LNM, n° 1829, Springer-Verlag, 2003.

[2] K. ATIF, B. PLATEAU. *Stochastic Automata Network for modeling parallel systems*, in "IEEE Transactions on Software Engineering", vol. 17, n° 10, October 1991.

[3] B. GAUJAL, S. HAAR, J. MAIRESSE. *Blocking a Transition in a Free Choice Net, and what it tells about its throughput*, in "Journal of Computer and System Sciences", vol. 66, n° 3, 2003, p. 515-548.

[4] J.-M. VINCENT. *Some Ergodic Results on Stochastic Iterative Discrete Event Systems*, in "Discrete Event Dynamic Systems", vol. 7, n° 2, 1997, p. 209-232.

Year Publications

Doctoral Dissertations and Habilitation Theses

[5] J.-M. N'LONG 2. *Conception et réalisation d'un intergiciel pour la résilience d'applications parallèles distribuées sur un intranet et Internet*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, July 2008.

[6] L. NUSSBAUM. *Étude de systèmes distribués à grande échelle*, Ph. D. Thesis, Université Joseph Fourier, Grenoble, December 2008.

Articles in International Peer-Reviewed Journal

[7] O. BEAUMONT, L. CARTER, J. FERRANTE, A. LEGRAND, L. MARCHAL, Y. ROBERT. *Centralized Versus Distributed Schedulers for Multiple Bag-of-Tasks Applications*, in "IEEE Trans. Parallel Distributed Systems", vol. 19, n^{o} 5, May 2008, p. 698–709, http://mescal.imag.fr/membres/arnaud.legrand/articles/tpds_msma.pdf.

[8] M. BEN MAMOUN, J.-M. FOURNEAU, N. PEKERGIN. *Analyzing Weighted Round Robin policies with a Stochastic Comparison Approach*, in "Computers and Operation Research", vol. To appear, 2008.

[9] A. BOUILLARD, B. GAUJAL. *Backward Coupling in Bounded Free-Choice Nets Under Markovian and Non-Markovian Assumptions*, in "Journal of Discrete Event Dynamics Systems, theory and applications", Special issue of selected papers from the Valuetools conference, vol. 18, 2008, p. 473-498, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/JDEDS-08b(2).ps.

[10] A. BOUILLARD, B. GAUJAL, E. THIERRY, S. LAGRANGE. *Optimal end-to-end routing for networks with multiplexing*, in "Performance Evaluation", special issue on Valuetools selected papers, vol. 65, 2008, p. 883-906, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/PEVA-08.ps.

[11] J.-M. FOURNEAU, B. PLATEAU, W. J. STEWART. *An algebraic condition for product form in stochastic automata networks without synchronizations*, in "Performance Evaluation", vol. 85, 2008, p. 854-868.

[12] B. GAUJAL, J. MAIRESSE. *Minimization of circuit registers: retiming revisited*, in "Discrete Applied Mathematics", vol. 56, 2008, p. 3498- 3505, http://www-id.imag.fr/Laboratoire/Membres/Gaujal_Bruno/Publications/DAM7021.pdf.

[13] A. LEGRAND, A. SU, F. VIVIEN. *Minimizing the Stretch When Scheduling Flows of Divisible Requests*, in "Journal of Scheduling", 2008, http://www.springerlink.com/content/7004222772015672/.

[14] I. SBEITY, L. BRENNER, B. PLATEAU, W. J. STEWART. *Phase-type distributions in stochastic automata networks*, in "European Journal of Operational Research", vol. 186, n^{o} 3, 2008, p. 1008–1028.

[15] I. SZCZESNIAK, T. CZACHORSKI, J.-M. FOURNEAU. *Packet Loss Analysis in Optical Packet-Switched Networks with Limited Deflection Routing*, in "Photonic Network Communications", vol. To appear, 2008.

[16] M. TCHUENTÉ, P. M. YONTA, J.-M. NLONG II, Y. DENNEULIN. *On the Minimum Average Distance Spanning Tree of the Hypercube*, in "Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications", 2008.

Articles in National Peer-Reviewed Journal

[17] M. MARTINASSO, J.-F. MÉHAUT. *Modèles de communication sur grappes de calcul multiprocesseurs*, in "Techniques et Sciences Informatiques", vol. 26, 2008.

[18] L. NUSSBAUM, O. RICHARD. *Une Plate-forme d'Émulation Légère pour Étudier les Systèmes Pair-à-Pair*, in "Technique et Science Informatiques - numéro spécial RenPar'17", vol. 26, 2008.

[19] C. PLUMEJEAUD, J.-M. VINCENT, C. GRASLAND, J. GENSEL, H. MATHIAN, S. GUELTON, J. BOULIER. *HyperSmooth : calcul et visualisation de cartes de potentiel interactives*, in "RNTI", 2008, to appear, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/RNTI-2008.pdf.

International Peer-Reviewed Conference/Proceedings

[20] N. ABED, S. TRIPAKIS, J.-M. VINCENT. *Resource-Aware Verification using Randomized Exploration of Large State Spaces*, in "SPIN, Los Angeles", Jun 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/Abed-Vincent-SPIN-2008.pdf.

[21] A. ANDRZEJAK, D. KONDO, D. P. ANDERSON. *Ensuring Collective Availability in Volatile Resource Pools Via Forecasting*, in "DSOM", 2008, p. 149-161, http://mescal.imag.fr/membres/derrick.kondo/pubs/andrzejak_dsom08.pdf.

[22] F. ARAUJO, P. DOMINGUES, D. KONDO, L. M. SILVA. *Using Cliques of Nodes to Store Desktop Grid Checkpoints*, in "Coregrid Integration Workshop, Crete, Greece", April 2008, http://mescal.imag.fr/membres/derrick.kondo/pubs/araujo_coregrid08.pdf.

[23] R. BERTIN, A. LEGRAND, C. TOUATI. *Toward a Fully Decentralized Algorithm for Multiple Bag-of-tasks Application Scheduling on Grids*, in "IEEE/ACM International Conference on Grid Computing (Grid), Tsukuba, Japan", 2008.

[24] L. BRENNER, P. FERNANDES, J.-M. FOURNEAU, B. PLATEAU. *Modelling Grid5000 point availability with SAN*, in "3rd International Workshop on Practical Applications of Stochastic Modelling (PASM 2008), Palma de Mallorca, Spain", sep 2008, p. 0–14.

[25] A. BUŠIĆ, J.-M. FOURNEAU. *Stochastic Bounds for Partially generated Markov chains: an algebraic approach*, in "Fith European Performance Engineering Workshop, EPEW 2008, Proceedings", Lecture Notes in Computer Science, Springer, 2008.

[26] A. BUŠIĆ, B. GAUJAL, J.-M. VINCENT. *Perfect Simulation and Non-monotone Markovian Systems*, in "3rd International Conference Valuetools'08, Athens, Greece", ICST, October 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/Ana-envelopes-2008.pdf.

[27] H. CASANOVA, A. LEGRAND, M. QUINSON. *SimGrid: a Generic Framework for Large-Scale Distributed Experiments*, in "Proceedings of the 10th Conference on Computer Modeling and Simulation (EuroSim'08)", 2008.

[28] P. COUCHENEY, C. TOUATI. *Replicator Dynamics Based Adaptive Algorithm for Heterogeneous Wireless Systems*, in "Proceedings of the 13th International Symposium on Dynamic Games and Applications (ISDG'2008)", 2008.

[29] P. FERNANDES, J.-M. VINCENT, T. WEBBER. *Perfect Simulation of Stochastic Automata Networks*, in "Analytical and Stochastic Modeling Techniques and Applications, 15th International Conference, ASMTA 2008, Nicosia, Cyprus, Proceedings, Nicosia, Cyprus", K. AL-BEGAIN, A. HEINDL, M. TELEK (editors), Lecture Notes in Computer Science, vol. 5055, Springer, Jun 2008, p. 249–263, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/ASMTA-2008.pdf.

[30] J.-M. FOURNEAU. *Discrete Time Markov chains competing over resources: product form steady-state distribution*, in "Fith International Conference on the Quantitative Evaluation of Systems (QEST 2008), St Malo, France", IEEE Computer Society, 2008.

[31] J.-M. FOURNEAU. *Multiclass G-Networks of Processor Sharing Queues with Resets*, in "Analytical and Stochastic Modeling Techniques and Applications, 15th International Conference, ASMTA 2008, Nicosia, Cyprus, Proceedings", K. AL-BEGAIN, A. HEINDL, M. TELEK (editors), Lecture Notes in Computer Science, vol. 5055, Springer, 2008, p. 221-233.

[32] J.-M. FOURNEAU. *Product form steady-state distribution for Stochastic Automata Networks with Domino Synchronizations*, in "Fith European Performance Engineering Workshop, EPEW 2008, Proceedings", Lecture Notes in Computer Science, Springer, 2008.

[33] A. HARBAOUI, B. DILLENSEGER, J.-M. VINCENT. *Performance characterization of black boxes with self-controlled load injection for simulation-based sizing*, in "CFSE, Fribourg", Feb 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/CFSE-2008.pdf.

[34] H. JOUMAA, Y. DEMAZEAU, J.-M. VINCENT. *Evaluation of Multi-Agent Systems: The case of Interaction*, in "ICTTA, Damas, Syria", Apr 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/ICTTA-2008.pdf.

[35] D. KONDO, A. ANDRZEJAK, D. P. ANDERSON. *On Correlated Availability in Internet Distributed Systems*, in "IEEE/ACM International Conference on Grid Computing (Grid), Tsukuba, Japan", 2008, http://mescal.imag.fr/membres/derrick.kondo/pubs/kondo_grid08.pdf.

[36] C. PLUMEJEAUD, J.-M. VINCENT, C. GRASLAND, S. BIMONTE, H. MATHIAN, S. GUELTON, J. BOULIER, J. GENSEL. *A System for Interactive Spatial Analysis via Potential Maps*, in "W2GIS, Shanghai, China", November 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/W2GIS-2008.pdf.

[37] C. POUSA, V. MARANGOZOVA-MARTIN, J.-F. MÉHAUT, F. DUPROS, A. CARISSIMI. *Explorando Afinidade de Memória em Arquiteturas NUMA*, in "Proceedings of IX Simpósio em Sistemas Computacionais (WSCAD-SSC 2008), Campo Grande, Brazil", October 2008.

[38] B. VIDEAU, O. RICHARD. *Expo : un moteur de conduite d'expériences pour plates-formes Dédiées*, in "Actes de CFSE'6", 2008.

[39] J. VIENNE, M. MARTINASSO, J.-M. VINCENT, J.-F. MÉHAUT. *Predictive models for bandwidth sharing in high performance clusters*, in "Proceedings of the IEEE Cluster Conference, Tsukuba, Japan", September 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/Cluster-2008.pdf.

[40] J.-M. VINCENT. *Perfect sampling of stationary rewards of Markov chains*, in "IWAP, Compiègne, France", Jul 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/IWAP-2008.pdf.

[41] J.-M. VINCENT. *Perfect simulation, monotonicity and finite queueing networks*, in "QEST, Saint-Malo", Sep 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/Vincent-QEST-2008.pdf.

[42] B. YENKÉ, J.-F. MÉHAUT, M. TCHUENTÉ. *Scheduling deadline constrained checkpointing on virtual clusters*, in "Proceedings of the IEEE Asia-Pacific Services Computing Conference (APSCC), Yilan, Taiwan", December 2008.

National Peer-Reviewed Conference/Proceedings

[43] H. ADAMOU, J.-F. MÉHAUT. *Performance des stratégies de répartition des tâches et des données sur grille: Impact du partage de données*, in "Proceedings du Colloque Africain de Recherche en Informatique (CARI 08), Rabat, Maroc", October 2008.

[44] V. BERTEN, A. BUŠIĆ, B. GAUJAL, J.-M. VINCENT. *Can we use perfect simulation for non-monotonic Markovian systems ?*, in "ROADEF, Clermont-Ferrand", February 2008, http://www-id.imag.fr/Laboratoire/Membres/Vincent_Jean-Marc/papers/Roadef-2008.pdf.

[45] K. GEORGIEV, L. GENOVESE, T. DEUTSCH. *Analyse et Evaluation d'un code de nanosimulation ab initio sur architectures multiprocesseurs*, in "Proceedings des Rencontres Francophones du Parallélisme, RenPar'18, Fribourg, Suisse", February 2008.

[46] E. SAULE, B. VIDEAU. *PaSTeL. Une implantation parallèle de la STL pour les architectures multi-coeurs : une analyse des performances*, in "Proceedings des Rencontres Francophones du Parallélisme, RenPar'18, Fribourg, Suisse", February 2008, http://www-id.imag.fr/~saule/papers/renpar08-SV.pdf.

[47] B. YENKÉ. *Prédiction des performances des opérations de sauvegarde/reprise sur cluster virtuel*, in "Proceedings des Rencontres Francophones du Parallélisme, RenPar18, Fribourg, Suisse", February 2008.

Scientific Books (or Scientific Book chapters)

[48] H. CASANOVA, A. LEGRAND, Y. ROBERT. *Parallel Algorithms*, Chapman & Hall, 2008, http://www.crcpress.com/shopping_cart/products/product_detail.asp?sku=C9454&isbn=9781584889458&parent_id=1153&pc=.

Research Reports

[49] R. BERTIN, A. LEGRAND, C. TOUATI. *Toward a Fully Decentralized Algorithm for Multiple Bag-of-tasks Application Scheduling on Grids*, Research Report, n^{o} 6537, INRIA, May 2008, https://hal.inria.fr/inria-00279993.

[50] P. COUCHENEY, B. GAUJAL, C. TOUATI. *A Distributed Algorithm for Fair and Efficient User-Network Association in Multi-Technology Wireless Networks*, Technical report, n^{o} RR-6653, INRIA, 2008.

[51] P. COUCHENEY, C. TOUATI, B. GAUJAL. *A Distributed Algorithm for Fair and Efficient User-Network Association in Multi-Technology Wireless Networks*, Research Report, n^{o} 6653, INRIA, September 2008, https://hal.inria.fr/inria-00322403.

[52] N. GAST, B. GAUJAL. *Distributing labels on infinite trees*, Technical report, n^{o} RR-6630, INRIA, 2008, http://hal.inria.fr/docs/00/31/88/72/PDF/RR-6630.pdf.

Other Publications

[53] D. BARTH, J.-M. FOURNEAU, D. NOTT. *Two cycles routing and End to End delay bound in all optical Network*, 2008, Photonics in Switching 2008, Poster Session.

References in notes

[54] A. LEBRE, Y. DENNEULIN. *aIOLi: An Input/Output LIbrary for cluster of SMP*, in "Proceedings of CCGrid 2005, Cardiff, Pays de Galles", 2005.